OVERVIEW OF THE MATCH COMPILER FOR COMPILING MATLAB PROGRAMS INTO HARDWARE

Prith Banerjee, Malay Haldar, Anshuman Nayak and Alok Choudhary
Electrical and Computer Engineering, Northwestern University
2145 Sheridan Road, Evanston, IL-60208
banerjee@ece.northwestern.edu

ABSTRACT

Efficient high-level design tools that can map behavioral descriptions of signal and image processing applications to FPGA architectures are one of the key requirements to fully leverage FPGAs for high-throughput computations and meet time to market pressures. Currently, most FPGA designs are entered at the level of Register Transfer Level (RTL) VHDL or Verilog. It is widely recognized that there is a need for design tools at the high level using languages such as C/C++ or MATLAB. MATLAB is an extremely popular language in the signal and image processing community with over 500,000 users. A direct synthesis path from MATLAB into hardware would be very useful. The MATCH compiler at Northwestern University [1] takes as input algorithms described in MATLAB, and generates Register Transfer Level (RTL) VHDL. The RTL VHDL then can be mapped to FPGAs using commercial tools. The input application is mapped to multiple FPGAs by parallelizing the application and embedding computation and synchronization primitives automatically. Our compiler infers the minimum number of bits required to represent the variables through a precision inferencing analysis framework. The compiler can leverage optimized Intellectual Property (IP) cores to enhance the hardware generated. The compiler also exploits parallelism in the input algorithm by pipelining in the presence of resource constraints. We demonstrate the utility of the compiler by synthesizing hardware for a couple of signal/image processing algorithms and comparing them to manually designed hardware.

1. INTRODUCTION

The concept of using FPGAs for custom computing evolved in the late 1980's. Wide adoption of the concept, however, has gained grounds only recently. One of the principal enabling factor was the availability of commercial synthesis and physical placement tools that raised the level of design abstraction to hardware description languages such as VHDL/Verilog. With gate counts for modern FPGAs reaching millions, we are poised for yet another revolution. The goal this time is to raise the level of abstraction to general purpose programming languages such as C/C++, Java and MATLAB Current design methodologies rely on the expertise of the hardware engineer to map the application onto a FPGA board. While this enables a lot of flexibility and fine grained control over the design, it also introduces a lot of logic design at a very low level. Not only is this process tedious and error-prone requiring costly debugging iterations, much of the work can be automated resulting in designs which are correct by construction. Another key aspect of mapping applications onto hardware is to exploit coarse and fine grained parallelism in the application. Again, concurrent simulations of multiple states is not the easiest thing to manage. Many mature and advanced compiler techniques exist that can discover and exploit parallelism and weigh different trade-offs automatically. All this will relieve the designer to focus on high level algorithmic aspects rather than learning about new board architectures or ways to boost performance by low level manipulations.

Synthesizing hardware from general purpose languages has received attention in both industry and academia. A broad classification of the different approached is possible from two perspectives :

1. *Target Language* : The approaches can be categorized according to the languages they target for synthesis. C/C++ has been the most popular choice [5, 8, 9, 10, 11, 13, 14, 15, 18, 12]. Java has been the focus of some recent works [17, 16]. Our focus is on MATLAB.
2. *Parallelism Specification* : The approaches can be classified depending on whether they attempt to automatically parallelize the input applications [15, 18, 13, 12] or they depend on the user to specify the parallelism [5, 11, 9, 10, 14, 17]. Depending on the user to specify the parallelism simplifies the compiler a lot, but it typically requires modifications/additions to the target language. It also burdens the user to extract the parallelism. User specified parallelism approaches does raise the design abstraction from VHDL/Verilog, but still require considerable manual iterations and interventions. Automatic parallelization makes the compiler complex but it doesn't require any modifications to the language and the user is not burdened with finding parallelism. This cuts down manual iterations to a minimum. Our approach is automatic parallelization, but experienced users can also direct the compiler through directives.

Optimized hardware synthesis from a general purpose language is a very complex task and the associated compiler framework has many components. The components include the front-end of the compiler dealing directly with the target language, the intermediate synthesis framework, the optimizations involved in synthesizing the hardware and the back-end which outputs the hardware and interfaces with lower level tools. All the components have their own specific issues, which were addressed in individual bits and pieces with many alternative solutions [1, 3, 4, 8, 11, 12, 13, 15, 14, 17, 10, 9, 18]. Our attempt in this paper is to discuss the complete system of an optimizing hardware synthesis tool and put forward a working combination of the mass of solutions contributed for each aspect of the compiler.

2. USE OF MATLAB FOR SYNTHESIS

While most of the industry and academia has focused on C/C++ as the system description language, our main focus is on MATLAB® [2]. Whereas many synthesis issues seem independent of the input specification language, MATLAB does offer distinctive advantage due to the following two reasons :

1. MATLAB is extremely popular in the signal/image processing community with over $500,000$ users. MATLAB is more intuitive than C/C++ and it enables simulation and visualization of algorithms with much less effort than C/C++. A direct synthesis path from MATLAB without first converting it into

[1] This research was supported in part by DARPA under contract F30602-98-2-0144 and in part by NASA under contract NAS5-00212.

another language like C/C++ will be very useful and it will enable very rapid and easy evaluation of a lot of algorithms. Thus a designer will be able to directly see the tradeoffs resulting from high level algorithmic changes.

2. A key technique that enables multi-million gate designs is reuse of Intellectual Property (IP) cores. Such cores correspond to common functions such as FFT, Viterbi decoders and Matrix Multiplication. These functions are available in MATLAB as standard function calls and operators with standard interfaces. This feature becomes particularly useful in recognizing that a particular IP block can be used for part of the input application and how to generate the interface signals corresponding to it. In languages without standard library calls for the algorithms, there may be innumerable ways to specify and invoke the algorithms. In such a situation it becomes very difficult to recognize that an IP block can be used for part of the algorithm and generating the interfaces for it.

However, MATLAB does have some disadvantages. The two main issues in that respect are :

1. MATLAB doesn't have any notion of type/shape for its variables. This becomes a nightmare from the compiler perspective and using existing techniques, in most cases inefficient code is generated that covers all or many of the possible types/shapes of the variables. We have developed a type/shape algebra framework that enables accurate inferencing, leading to efficient hardware generation [1]. In spite of the inferencing framework, if the compiler is unable to do a satisfactory job, the user can force the type/shape of a variable through directives.
2. Simulation of scalarized MATLAB code is slower than a compiled approach. This is because MATLAB is an interpreted language which incurs a lot of overhead if simple computations are done in a loop. However, our focus is on signal/image processing kind of applications where arbitrary loops and array manipulations is not the norm. Regular loops with extensive use of library functions is more common for such applications for which MATLAB is ideally suited.

3. OVERVIEW OF COMPILATION PROCESS

We now present an overview of our compiler architecture. Figure 1 shows the different compiler phases. The front-end parses the input MATLAB program and builds a MATLAB AST(Abstract Syntax Tree). The input code may contain directives [1] regarding the types, shapes and precision of arrays that cannot be inferred, which are attached to the AST nodes as annotations. This is followed by a type-shape inference phase. MATLAB variables have no notion of type or shape. The type-shape phase analyzes the input program to infer the type and shape of the variables present for which type/shape is not provided by directives. This is followed by a scalarization phase where the operation on matrices are expanded out into loops. In case optimized library functions are available for a particular operation, it is not scalarized and the IP core corresponding to the library function is used instead. The scalarized code is then passed through the parallelization phase. The parallelization phase attempts to exploit coarse grain parallelism by either splitting a loop onto multiple FPGAs on the board (data-parallel approach) or by putting different tasks onto different FPGAs and pipelining the output of one to the input of another (systolic approach). The parallelization phase relies on communication libraries implemented for the target architecture board to communicate between the different FPGAs. A state machine description in VHDL is then synthesized from the parallelized scalarized MATLAB code for each of the FPGAs . Most of the hardware related optimizations are performed on the VHDL AST. A precision inference scheme finds the minimum number of bits required to represent each variable in the AST. The precision information is used in instantiating customized IP blocks corresponding to the functions and operators. Transformations are then performed on the AST to optimize it according to the memory accesses present in the program and characteristics of the external memory. This is followed by a phase to perform optimizations like pipelining under resource constraints that alter parts of the state machine that was constructed earlier. Finally a traversal of the optimized VHDL AST produces the output code.

4. EXPERIMENTAL SETUP AND BENCHMARKS

Our compiler is designed to produce code for most current FPGA architectures. The results presented in this paper are for hardware generated for the $WildChild^{TM}$ FPGA board from Annapolis Micro Systems. It is a VME compatible board with eight $Xilinx$ 4010 FPGAs and one $Xilinx$ 4028 FPGA. The $Xilinx$ 4028 has an external memory that is 32-bit wide with 2^{18} addressable locations. The memories connected to the 4010s are 16-bit wide.

The benchmarks include Matrix Multiplication, FIR filter, IIR filter, Sobel edge detection algorithm, an Average filter and a Motion Estimation algorithm. These benchmarks represent typical signal/image processing applications that are of interest to us. On one hand, such applications are important as they are representative of a class of applications that are predicted to be ubiquitous in next generation computing platforms, in environments that demand high throughput. On the other hand, these applications have inherent parallelism suitable for exploitation by implementation in customized hardware.

5. COMPILING MATLAB TO VHDL

One of the challenges in generating hardware from MATLAB is to figure out the type/shape of the variables. As shown in Figure 2, the semantics of an operator can depend on the assignments to the operands. To generate hardware, the compiler must figure

```
a = 1 ;          a = rand( 256 , 234 );
b = 3 ;          b = ones(234, 512 );

c = a * b ;      c = a * b ;

  ( i )            ( ii )
```

Figure 2. The semantics of an operator depends on the type/shape of the operands in MATLAB. In (i) $*$ is a scalar multiplication whereas in (ii) it is a matrix multiplication.

out the exact data type i.e, integer or floating point, or complex numbers etc. The compiler also needs to figure out the shape i.e, how many dimensions the matrix (array) has, what are the extents in each dimension, etc. Our type shape algebra framework automatically figures out the type-shape of the variables [1]. In pathological cases where the compiler is unable to infer the type/shape of the variables, the user can assist the compiler by specifying the type/shape of selected variables. Once the type/shape of the variables are determined, the matrix operations are scalarized, the operations are expanded out into loops. Scalarization of the MATLAB AST is necessary when the objective is to perform a source-to-source transformation to a target language that is statically typed and which only supports elemental operations. MATLAB is an array-based language with many built-in functions to support array operations. Hence, to generate a VHDL description, it is necessary that the corresponding MATLAB AST is scalarized. Figure 3 shows an example where VHDL code is generated corresponding to a matrix multiply operation. Extensive discussion of VHDL generation from MATLAB is reported in [1]. The framework is capable of handling multi-dimension matrices which are mapped to an external memory. In addition, the loop and function call constructs of MATLAB are also supported. Figure 4 shows the experimental results of execution times of the benchmarks on a $Xilinx$ 4028 using manual and compiler approaches. As can be seen, the manually designed hardware on the average is five times better than the compiler output, noting that it took several months to complete the manual designs while the compiler generated the hardware in a

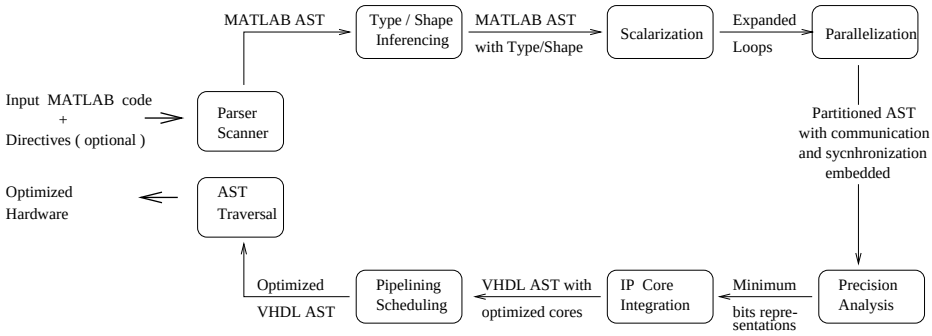

Figure 1. Overview of the synthesis framework.

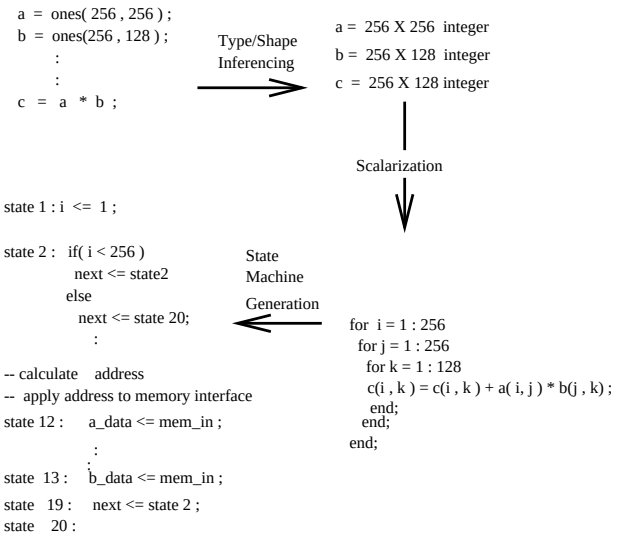

Figure 3. An example showing how a state machine is synthesized for matrix multiplication by first doing type/shape analysis, followed by scalarization.

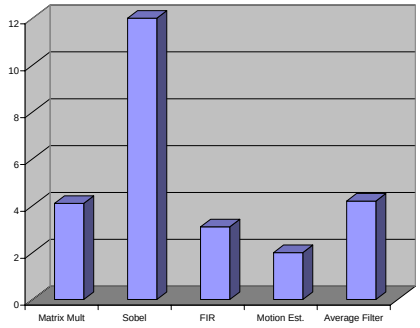

Figure 4. Ratio of execution times of compiler generated hardware compared to manually designed hardware is shown. For example, for the matrix multiplication benchmark, the compiler generated hardware is 4 times slower than the manually designed hardware.

matter of minutes.**Reduction of design time is the key advantage of using the compiler**. In the next few sections, we elaborate how our compiler closes the performance gap between its output and the manually designed hardware.

6. PRECISION INFERENCING

One important factor in generating customized hardware for an application is to efficiently utilize the silicon budget available. A key observation in this regard is that most image/signal processing computations are confined to $8 \backsim 16$ bits. To fully leverage this fact, the minimum number of bits required to represent each variable must be inferred and appropriate operators instantiated in place of generic 32-bit operators. However, figuring out the precision manually in a real life design can be very tiresome and error prone. Our precision inferencing algorithm propagates value range information back and forth the AST to figure out the minimum bits required to represent a variable, see Figure 5. In case where

```
a = 8 ;       % 4 bits required
b = 4 ;       % 3 bits required
d = a + b ;   % 4 bits required
e = b + input()  % unknown , give
                 % directive
```

Figure 5. Illustration of precision inferencing.

the precision of variables cannot be determined statically, the user can specify the precision by a directive; otherwise the most conservative estimate is taken. For floating point variables, in association with the precision inferencing algorithm an error analysis and propagation scheme is included. The error analysis determines the resolution of the floating point variables needed given a specified error that can be tolerated at the output. Details of the precision and error analysis algorithms can be found in [1]. Figure 6 shows the savings of resources in terms of CLBs when the precision inferencing algorithm is applied as opposed to instantiating generic 32 bit operators.

7. PIPELINING

A close study of the manually designed hardware and the compiler generated hardware showed that the principal reason behind the better performance of the manually generated hardware was exploitation of fine grain parallelism and pipelining of the memory accesses. This prompted us to devise an automated way of pipelining the memory accesses and exploit fine grained parallelism. Our pipelining framework achieves this objective, an overview of which is given in Figure 7. The pipelining phase starts by performing dependency analysis of the loops and basic blocks. The GCD test is employed to figure out loop carried dependencies. In case there are no backward dependencies in a loop, the loop is deemed pipelinable. Next the number of memory ports are read as input to the pipelining algorithm. The pipelining algorithm then performs modulo scheduling [1] which overlaps different iterations of a loop such

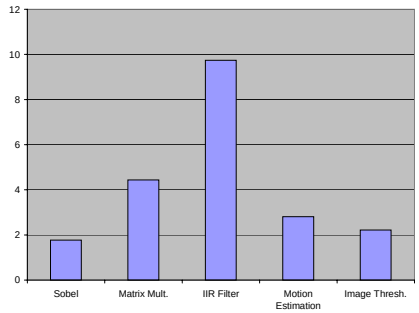

Figure 6. Ratio of the resource utilization in terms of CLBs while instantiating 32-bit operators as compared to determining the minimum number of bits required by precision inferencing.

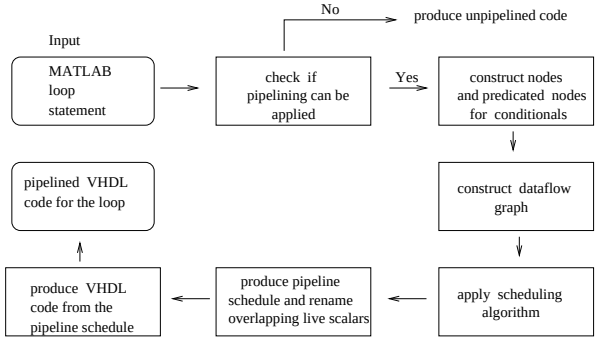

Figure 7. An overview of the pipelining framework.

that number of memory access in any state does not exceed the number of memory ports specified. The modulo scheduling algorithm can be based on either ASAP (as soon as possible) or ALAP (as late as possible) algorithms. The reason the pipelining algorithm is based on memory ports is that many of the image/signal processing applications are memory bound. They tend to perform simple operations on relatively large data sets that reside in external memories. Hence, optimizing the memory accesses in general has a huge impact on performance. Conflicts created in variables due to overlapping of iterations is solved by renaming the variables as discussed in [1]. Figure 8 shows the impact of pipelining on performance. The compiler generated pipelined hardware matches

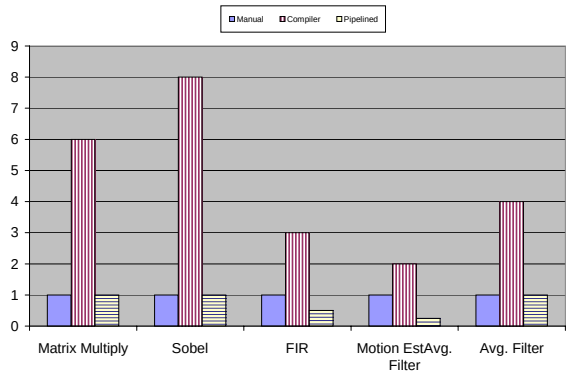

Figure 8. Ratio of the execution times of the compiler generated hardware with and without pipelining is shown, normalized to the execution time of manually designed hardware. For example, for the Sobel benchmark, the compiler generated hardware without pipelining is 8 times slower, whereas the pipelined hardware is as fast as the manual design.

the manual designs in most cases. In fact, the compiler generated pipelined hardware fares better than the manually designed hardware for the FIR and Motion Estimation benchmark. This is due to the fact that manual designers typically pipeline and exploit parallelism *within* a single iteration of the loop. The compiler can handle much more complexity and exploits parallelism *across* the different iterations of the loops. For example, the pipeline kernel synthesized for the motion estimation benchmark contained 200 concurrent statements spanning 5 iterations. Such complexity can only be handled in an automated fashion.

8. SUMMARY

We now present the result of performing all the optimizations together and compare with manually designed hardware. We would like to emphasize once more that the manually generated hardware took months of design effort whereas the compiler generated the hardware in a matter of minutes. While a massive reduction in design time is achieved, the quality of the hardware generated was not compromised. Indeed, the hardware generated by the compiler were very close to the manually generated hardware in performance, in fact better in some cases. Figure 9(i) shows an input image to the Sobel edge detection algorithm. Figure 9(ii) shows the output of the Sobel edge detection algorithm as simulated in the MATLAB interpretor.

The same MATLAB code was then used to synthesize a pipelined hardware. The output of the hardware is shown in Figure 9(iii). The output matches the simulation result pixel by pixel. The designs generated by the compiler are correct by construction and do not require debugging iterations. Figure 10 shows the comparison of the execution times of the compiler generated hardware with the optimizations against the manually designed hardware for the benchmarks. Figure 11 shows a comparison of the resource utilization for the same. The performance of the compiler output and manually optimized hardware are comparable. The resource utilization of the compiler generated hardware are within a factor of four of the manually designed hardware.

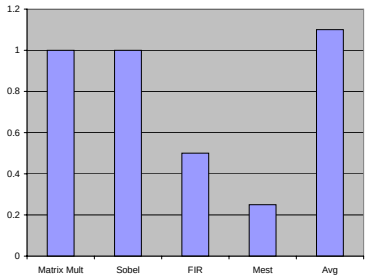

Figure 10. Ratio of the execution times of the compiler generated hardware with the optimizations as compared to the manually generated hardware.

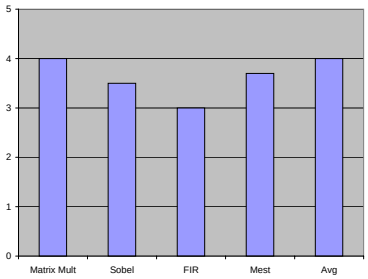

Figure 11. Ratio of the CLBs used by the compiler generated hardware with optimizations as compared to the manually generated hardware.

(i)Input Image (ii)MATLAB Interpreter (iii)Annapolis $Wildchild^{TM}$

Figure 9. A grayscale image is shown in (i). Output of the Sobel edge detection algorithm simulated in the MATLAB interpretor is shown in (ii). The MATLAB code is used to synthesize hardware for the Annapolis $Wildchild^{TM}$ board and its output is shown in (iii).

9. CONCLUSIONS AND FUTURE WORK

In conclusion we have presented a compiler capable of generating highly optimized hardware from applications described in MATLAB. A set of effective optimizations implemented in the compiler ensures that the quality of the output hardware is comparable to manually optimized hardware. The optimizations include parallelization, precision inferencing, IP core integration and pipelining. The effectiveness of the compiler was demonstrated by synthesizing hardware for a couple of signal/image processing applications. The outputs of the synthesized hardware were functionally verified against the outputs of the MATLAB interpretor. The execution times were almost equivalent to manually designed hardware, in fact superior in some cases were large amount of parallelism was available across loops. The resource utilization were within a factor of four of the manual designs. All this was achieved while reducing the design time from months to minutes.

The major focus of our current and future work is in the following two directions

1. We are investigating methods to identify and utilize opportunities to synthesize on-chip caches to reduce the memory traffic and boost the performance of the synthesized hardware.
2. We are concentrating on accurate prediction of the resource and routing resources needed for a particular design to achieve design closure in minimum iterations possible.

REFERENCES

[1] *Various Reports*, details masked.

[2] The Mathworks Homepage, www.mathworks.com.

[3] G. D. Micheli,*Synthesis and Optimization of Digital Circuits*,pg. 185-265, ISBN-0-07-016333-2, McGraw-Hill, Inc.

[4] D. Gajski, N. Dutt, A. Wu and S. Lin,*High-Level Synthesis: Introduction to Chip and System Design*, ISBN-0-7923-9194-2, Kluwer Academic Publishers.

[5] The SystemC Initiative, www.systemc.org.

[6] *Xilinx Core Solutions Databook, Second Edition*, Xilinx Inc.

[7] *Altera 1999 Intellectual Property Catalog*, Altera Inc.

[8] G. De Micheli, *Hardware Synthesis from C/C++ Models*, Proc. Design, Automation and Test in Europe Conference and Exhibition, March 1999.

[9] A. Ghosh, J. Kunkel and S. Liao, *Hardware Synthesis from C/C++*, Proc. Design, Automation and Test in Europe Conference and Exhibition, 1999.

[10] G. Arnout, *C for System Level Design*, Proc. Design, Automation and Test in Europe Conference and Exhibition, March 1999.

[11] J. Hammes, B. Rinker, W. Bohm and W. Najjar, *Cameron: High Level Language Compilation for Reconfigurable Systems*, Proc. Parallel Architectures and Compilation Techniques (PACT'99), October 1999.

[12] J. Babb, M. Rinard, C.A. Moritz, W. Lee, M. Frank, R. Barua, S. Amarasinghe *Parallelizing Applications into Silicon*, FCCM 1999.

[13] M. Weinhardt and W. Luk, *Pipeline Vectorization for Reconfigurable Systems* , Proc. Field-Programmable Custom Computing Machines, April 2000.

[14] M. Gokhale, J. Stone, J. Arnold and M. Kalinowski, *Stream-Oriented FPGA Computing in the Streams-C High Level Language*, Proc. Field-Programmable Custom Computing Machines, April 2000.

[15] Y. Li, T. Callahan, E. Darnel, R. Harr, U. Kurkure and J. Stockwood, *Hardware-Software Co-Design of Embedded Reconfigurable Architectures*, Proc. 37th DAC, June 2000.

[16] R. Helaihel and K. Olukotun, *Java as a Specification Language for Hardware-Software Systems*, Proc. International Conference on Computer-Aided Design, pp. 690-697. November 1997.

[17] B. L. Hutchings and B. E. Nelson,*Using General-Purpose Programming Languages for FPGA Design*, Proc. 37th Design Automation Conference, June 2000.

[18] C Level Design, Inc.,*System Compiler : Compiling ANSI C/C++ to Synthesis-ready HDL* , http://www.cleveldesign.com .